American Journal of Theoretical and Applied Statistics
2015; 4(3): 99-111
Published online April 3, 2015 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20150403.15
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Application of a Bivariate Poisson Model in Devising a Profitable Betting Strategy of the Zimbabwe Premier Soccer League Match Results

Desmond Mwembe, Lizwe Sibanda, Ndava Constantine Mupondo

National University of Science and Technology, Department of Statistics and Operations Research, Ascot, Bulawayo, Zimbabwe

Email address:
desmwembe@gmail.com (D. Mwembe), maquaqua@gmail.com (L. Sibanda), mupondocn@gmail.com (N. C. Mupondo)

To cite this article:
Desmond Mwembe, Lizwe Sibanda, Ndava Constantine Mupondo. Application of a Bivariate Poisson Model in Devising a Profitable Betting Strategy of the Zimbabwe Premier Soccer League Match Results. *American Journal of Theoretical and Applied Statistics.* Vol. 4, No. 3, 2015, pp. 99-111. doi: 10.11648/j.ajtas.20150403.15

Abstract: The study seeks to construct a profitable betting strategy for soccer results by developing a bivariate Poisson model for the analysis and computation of probabilities for football match outcomes. The dependence coefficient is estimated from Monte Carlo simulation and the scoring intensities are estimated from a log-linear model. The hypothesis tests show that the home-ground effect exists for some, but not all teams in the Zimbabwe Premier Soccer League. The profitable betting rule is to place a bet on the outcome of a particular match when a model's probabilistic forecast suggests a sufficient edge over the bookmaker's implied probability.

Keywords: Betting Strategy, Soccer, Home-Ground Advantage, Scoring Intensities, Fixed-Odds

1. Introduction

The gambling industry has grown and evolved substantially over the years with people betting on lotto, casino, scratch cards and sports, with soccer betting topping the sport betting list. Gambling in Zimbabwe is legal through the Lotteries and Gaming Act [Chapter10:26] of 2000.

This study will focus on modelling, statistically, the match results in the Zimbabwe Premier Soccer League so as to devise a profitable betting strategy. The Zimbabwe Premier Soccer League (ZPSL) is the top professional division of the Zimbabwe Football Association (ZIFA 2013). The League consists of 16 teams that play a total of 30 matches and the season runs from March to November in a year.

When betting on the result of a soccer match, the gambler may choose one out of the three different outcomes. However, the probabilities for all of the possible outcomes are not exactly known, so the betting house comes up with its own estimates. The odds prices are then fixed according to those estimates, which are slightly adjusted to provide the house edge.

1.1. Statement of the Problem

When placing bets, some gamblers are attracted to betting on so-called 'certainties' irrespective of the odds on offer and others may bet on the football team thatthey support to win out of loyalty. Gamblers often lose large sums of money or gain insignificant profits because of their uninformed decisions when placing bets. Forest and Simmons (2001) also highlight those odds compilers use several *ad hoc* techniques and use their opinion in compiling final prices. A lot of probability estimates are required in compiling the odds thus without the use of a proper statistical tool, it is possible that the bookmaker might make inaccurate probability estimates for some markets thereby over-pricing or under-pricing the odds.

1.2. Aim

Create a model that is capable of predicting results of football matches with reasonable accuracy, thereby creating a betting strategy with positive expected returns.

1.3. Research Objectives

The objectives of this study were to determine whether the home ground advantage exists in the Zimbabwe Premier Soccer league, to estimate the home ground advantage for

each team, if it exists, to estimate the dependence coefficient between the home and away goals, to estimate the attack and defence capabilities for the teams, to estimate the goal scoring intensities for soccer teams, to compute probabilities for match outcomes using a statistical model and consequently devise a profitable betting strategy.

2. Literature Review

For one to devise a profitable betting strategy, there is need to accurately predict the outcome of a soccer match between two teams playing against each other. Cengiz. *et al.*(2012), Goddard and Asimakopoulos (2004) and Dixon and Coles (1997)cite Moroney (1956) as the first researcher to propose a statistical model for soccer predictions. Moroney (1956) is said to have examined the distribution of the number of goals scored in a match and suggested that the Poisson distribution can be used to reasonably fit the results of soccer games and improvements could be obtained by working with the negative binomial distribution. This came from the assumption of a completely random process with fixed probability and team specific scoring probability.

Koopman and Lit (2012) stated that the number of goals scored by a team may depend on the attack capability of the team, the defence capability of the opposing team, the home ground advantage, if applicable, and the development of the match itself. A bivariate Poisson distribution is used in combination with a dependence parameter which allows for correlation between home and away scores.

Karlis and Ntzoufras (2003) highlighted in their study of the existence of a relatively low correlation between the numbers of goals scored by the two opponents. These were previously discussed by Maher (1982) as cited by Koopman and Lit (2012) and Dixon and Robinson (1998). They pointed out that this correlation has been ignored inmost modelling approaches since it demands more sophisticated techniques. The authors then advocated for the use of a bivariate Poisson distribution. Crowder *et al.*(2002) discussed dynamic models and developed a procedure for updating a team's scoring and goal conceding strength intensity parameters by analysing English Premier League matches played during the period 1992 to 1997. The procedure was claimed to be less computationally demanding than the one used by Dixon and Coles (1997) and Rue and Salvesen (1998).

Saavedra *et al.*(2012) analysed the home advantage of the first division of Spanish soccer from 1928 until 2011. The sample is of 80 seasons and 22015 games of the highest level in Spain. Their study shows that the advantage of playing at home exists and is significant. Oberhofer*et al* (2009) analysed the role of distance on professional team performance in the German Football League. They argued that a sports team might be less successful if the playing venue is relatively far away from the home location. Their findings suggested that distance exerts a significantly negative and non-monotonic impact on a soccer team's defensive performance, indicating that the performance of an away team becomes worse up to a certain distance.

Pollard (2006) outlined that the precise causes of home advantage in soccer and the way in which they affect performance are still not clear. The researcher gave a comprehensive review under the main hypothesized explanations for home advantage in soccer and these are: crowd effects, travel effects, familiarity, referee bias, territoriality, specific tactics, rule factors and psychological factors.

Their interactions and other factors that need to be taken into account when investigating home advantage were considered. The author pointed out that home advantage has been in existence at least since the start of organized football at the end of the 19th century and it is a worldwide phenomenon, but varies considerably from country to country.

Boyko*et al* (2007) investigated whether referee bias contributes to home advantage in the English Premiership football. An ordinal regression model was developed to determine whether various measures of home advantage are affected by the official for the match and by crowd size while controlling for team ability.

The study shows that sports with subjective officiating tend to experience greater home advantage and that referees' decisions can be influenced by crowd noise. The results of the study also confirm that referees are responsible for some of the observed home advantage in the English Premier League and suggest that home advantage is dependent on the subjective decisions of referees that vary between individuals.

Clarke and Norman (1995) used a range of non-parametric techniques to identify the effect of home advantage on match results. Home advantages for all teams in the English Football league from 1981/1982 to 1990/1991 are calculated and some reasons for their differences investigated. A paired home advantage is defined and has shown to be linearly related to the distance between club grounds. However, their study shows no evidence of higher home advantage for teams new to a league.

O`Shaughnessy (2012) derived the criterion one should use when investing in a 'win-draw-loss' betting market, with the important feature that profits are significantly higher by combining back and lay bets than by relying on one or the other.

O`Shaughnessy (2012) solved the Kelly criterion for the case of win-draw-loss markets. The researcher cited the Kelly (1956)'s criterion as a vital tool for both portfolio investors and gamblers. By maximising logarithmic utility and simultaneously minimising the risk of ruin, Kelly (1956) provided the formula that gamblers with perfect probabilistic knowledge must use to grow their bank at the largest expected rate.

Demir (2011a) explored the behavioural pattern of bettors on their betting choices for draws by using the FIFA 2010 World Cup South Africa. The World Cup Data showed that there is a draw bias among bettors as they prefer to bet mostly on win of a side. Demir (2011b) implemented the Fibonacci sequence on draws as a betting rule for 8 European soccer leagues for the seasons from 2005/2006 to 2008/2009.

The Fibonacci strategy for 8 soccer leagues of Europe for

4 seasons yielded positive returns for all cases and also controlling with simulated data the strategy was found to be in most circumstances profitable. The results indicated that the bookmakers are inefficient in terms of predicting the draws.

Demir (2011b) proposed that the gamblers could exploit this inefficiency by following Fibonacci strategy assuming they have enough financial liquidity. The capital needed to pursue this strategy was calculated, resorting to the Value at Risk (VaR) methodology. The VaR was revealed to be $143 (assuming that the first bet is $1) at 95% confidence level.

Vlastakis*et al* (2007) examined the potential for generating positive returns from wagering on soccer matches. They employed an arbitrage and a simple betting strategy based on a logit regression forecasting model. The research suggested that the differences in the bookmaker's odds can lead to profitable arbitrage opportunities. The authors showed that a betting strategy based solely on the information embedded in the bookmakers' odds can yield positive expected return.

Olesen (2008) evaluated different approaches individually to determine which strategy is better between the value betting and the threshold strategy. A data set containing an actual bookmaker's odds for several thousand matches were acquired.

The two betting strategies were examined. The expected value strategy was used together with three probability assessors; Gamblers' assessment, Poisson assessment and Dixon-Coles assessment. For both the value betting strategy and the threshold betting strategy, it was not possible to conclude that any of the two is able to produce a stable income on the leagues in question.

To beat the bookmakers as a gambler, Olesen (2008) noted that it is not only necessary to have better predictions, but it is also necessary that the predictions are significantly better in order to beat the margin the bookmaker gains by the theoretical payback percentage. The tests made in this study were made on only a half of a season.

The researcher acknowledged that this left a lot of things to chance. Perhaps this specific half season saw results which were very much different than the results usually are in the league therefore it is not possible to dismiss the betting strategies and the fact that it could be possible that they could generate a profit over a larger number of matches.

Goddard and Asimakopoulos (2004) claimed to have found a profitable betting strategy in their model. The model was used to test the efficiency of the prices quoted by bookmakers for fixed-odds betting on match results during four soccer seasons. Regression-based tests indicated that the model contained information about match outcomes that is not impounded into the bookmakers' prices. The bookmakers' prices were therefore concluded to be weak-form inefficient.

A strategy of selecting bets ranked in the top 15% by expected return according to the model's probabilities would have generated a positive return of at least 4% in each of the four seasons. The study showed some evidence that inefficiencies in the bookmakers' prices have diminished over time.

Martinnen (2001) investigated the possibility of creating a profitable betting strategy for soccer by building the Poisson model and examining its usefulness in the betting market. Comparisons of the Poisson model against other most commonly used prediction methods, such as Elo-ratings and multinomial ordered probit model were performed.

Forest and Simmons (2001) analysed patterns of odds in the British soccer betting market over four seasons from 1997. They discovered limited evidence for home-away and favourite-longshot bias identified in previous research.

Their study also considered the unexplored issue of the extent to which odds set by bookmakers reflect the different numbers of supporters that different football clubs have.

However, the growth of betting opportunities and greater competition had moved the market strongly towards efficiency. Dixon and Coles(1997) also investigated the inefficiencies in the UK betting market. A model built for soccer results predictions was used at each new time *t to* obtain current parameter estimates. Then, by comparing estimated result probabilities with bookmakers' odds, games which were advantageous to bet on were determined. The betting strategy proposed was to bet on all outcomes for which the ratio of model to bookmaker' probabilities exceeds a specified level.

3. Methodology

Due to relegations and promotions every season, soccer match results are collected for 10 seasons consecutively from 2004 to 2013.

3.1. Modelling Framework

The outcome of a match between the home football team *i* and the visiting football team *j* in week *t* is given by a pair of counts $(X; Y) = (X_{it}; Y_{jt})$, for $i \neq j = 1, \ldots, N$ and $t = 1, \ldots, n$, where n is the number of weeks available in our data set and N is the number of teams in the sample. The first count X_{it} is the number of goals scored by the home team i and the second count Y_{jt} is the number of goals scored by the visiting team j, in week t. Each pair of counts (X, Y) is assumed to be generated from the bivariate Poisson distribution with probability density function;

$$\mathrm{P}(\mathrm{X},\mathrm{Y};\ \lambda_x,\lambda_y,\gamma) = \exp\left(-\lambda_x-\lambda_y-\gamma\right)\frac{\lambda_x^{\mathrm{X}}}{\mathrm{X}!}\frac{\lambda_y^{\mathrm{Y}}}{\mathrm{Y}!}\sum_{k=0}^{\min(\mathrm{X},\mathrm{Y})}\binom{\mathrm{X}}{k}\binom{\mathrm{Y}}{k}k!\left(\frac{\gamma}{\lambda_x\lambda_y}\right)^k, \quad [1]$$

with λ_x and λ_y being scoring intensity coefficients for X and Y, respectively. Goal scoring intensities represents the number of goals the teams are expected to score against each other. The parameter γ is a coefficient that measures the dependence between the two counts X and Y. The means, variances and covariance for the home score X and the away score $Y$are given by,

$$E(X) = Var(X) = \lambda_x + \gamma, E(Y) = Var(Y) = \lambda_y + \gamma, Cov(X, Y) = \gamma \quad [2]$$

The correlation coefficient between X and Y is given by

$$\rho = \frac{\gamma}{\sqrt{(\lambda_x+\gamma)(\lambda_y+\gamma)}} \quad [3]$$

3.2. Home and Away Team Performance

A test on whether the home ground advantage really exists in the Zimbabwe Premier Soccer League is done. Hypotheses tests for a difference in means for two samples of size N with variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$ unknownwill be carried out to determine whether the mean number of home goals is significantly greater than the mean number of away goals in the league. Assuming that the populations are normally distributed, the variances are unknown and equal and the sample sizes do not exceed 40, therefore the hypotheses tests are based on the t-distribution.

The objective is to determine whether the home-ground performance of a team is better than the away-ground performance, hence we test the following hypotheses at $\alpha = 0.05$ level of significance;

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_0: \mu_1 - \mu_2 > 0$$

Where μ_1 and μ_2 are respective mean number of goals at home-ground and away grounds.

Let $X_{11}, X_{12}, \ldots, X_{1N}$ be a random sample of N observations from the total home-ground scores population and $X_{21}, X_{22}, \ldots, X_{2N}$ be a random sample of N observations from the away-ground scores population. Let $\bar{X}_1, \bar{X}_2, S_1^2$and S_2^2be the samplemeans and sample variances, respectively. The expected value and variance of the difference in sample means $\bar{X}_1 - \bar{X}_2$ is given by

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2; \qquad Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma^2}{N} + \frac{\sigma^2}{N} = \frac{2\sigma^2}{N} \quad [4]$$

So$\bar{X}_1 - \bar{X}_2$ is an unbiased estimator of the difference in means. Combining the twosample variancesS_1^2 and S_2^2 we form an estimator ofσ^2. The pooled estimator of σ^2, denoted by S_p^2 is defined by:

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \quad [5]$$

For the hypotheses tests, the test statistic is given by

$$t_0 = \frac{\bar{X}_1-\bar{X}_2}{S_p\sqrt{\frac{2}{N}}} \quad [6]$$

With *2N – 2* degrees of freedom. The null hypothesis $H_0$is rejected at the$\alpha = 0.05$ level of significance if the computed value of the test statistic$t_0 > t_{0.05,2N-2}$, the critical value. Rejecting the null hypothesis means that the home scores are significantly greater than away scores, thereby, leading to the conclusion that the home advantage exists.

3.3. Parameter Estimation

Using goals scored and conceded we estimate parameters to be used in the model. The parameters are; the home-ground advantage (if it exists), attack and defence capabilities, goal scoring intensities and the dependence coefficient between the two scores.

3.4. Home-Ground Advantage

Analysis of variance (ANOVA) procedure for testing the equality of several means was conducted. The equality of 40 treatment means$\mu_1, \ldots, \mu_{40}$, since there are 40 teams under study, is done. This is equivalent to testing the hypotheses

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_{40} = 0$$

against$H_1: \tau_i \neq 0$for at least one i.

When the null hypothesis is rejected in the ANOVA, then some of the treatment or factor level means are different. Fisher's Least Significant Difference (LSD) method has to be conducted to determine which means are significantly different.

Different home ground advantage coefficients for each team are estimated. Home-ground advantage δ_i for team *i*, $i = 1, \ldots, N$, was estimated by comparing the team's home and away scores as follows;

$$\delta_i = \frac{HD_i}{n_i} \quad [7]$$

where$HD_i$is the home-ground goal difference for team iand n_i is the total number of games played by team i.

3.5. Attack and Defence Capabilities

The attack and defence capabilities α_{it} and β_{it} are the numbers of goals expected to be scored and conceded respectively by a team. These are assumed to be time – varying and can be identified when we analyse football match results in a competition for a number of seasons as a time series panel of pairs of counts. The attack and defence capabilities are specified as p^{th}– order autoregressive processes. This leads to:

$$\alpha_{it} = \phi_{1,\alpha,i}\alpha_{i,t-1} + \phi_{2,\alpha,i}\alpha_{i,t-2} + \cdots + \phi_{p,\alpha,i}\alpha_{i,t-p} + \epsilon_{\alpha,it}$$

$$\beta_{it} = \phi_{1,\beta,i}\beta_{i,t-1} + \phi_{2,\beta,i}\beta_{i,t-2} + \cdots + \phi_{p,\beta,i}\beta_{i,t-p} + \epsilon_{\beta,it} \quad [8]$$

Where $\phi_{k,\alpha,i}$ and $\phi_{k,\beta,i}$,k= 1,2 are autoregressive coefficients and the disturbances $\epsilon_{\alpha,it}$ and $\epsilon_{\beta,it}$ are normally distributed error terms which are independent of each other for all $i = 1, \dots, N$ and all $t = 1, \dots, n$ .The processes are assumed to be independent of each other and stationary. It is required that $|\phi_{k,\alpha,i}| < 1$ and $|\phi_{k,\beta,i}| < 1$ for $k = 1,2$ and $i = 1, \dots, N$

3.6. Goal Scoring Intensities

The log-linear random effect model common in the statistical literature is used in modelling these parameters (Karlisand Ntzoufras 2003, 2009; Cengizet al.2012). In this study, another parameter, the goal ratings, is introduced. Let the goal scoring intensities λx and λy vary with the pairs of teams that play against each other. Furthermore, if these coefficients of goal scoring intensity change slowly over time, then the composition and the performance of the teams will change over time. The intensity of scoring for team i, when playing against team j, in week t is assumed to depend on the home-team goal rating, Gx,i, the away-team goal rating, Gy,j, the home-ground advantage,δ_i,the attack capability of team i,α_{it} and the defence capability of the opposing team j,β_{jt} for $i \neq j = 1, \dots, N$ and $t = 1, \dots, n$. The goal scoring intensities for home team i and away team j in week tare then given by;

$$\ln \lambda_{x,ijt} = G_{x,i}(d_{it}\delta_i + \alpha_{it} + \beta_{jt}) \ln \lambda_{y,ijt} = G_{y,i}(\alpha_{jt} + \beta_{it}) \quad [9]$$

where$i \neq j = 1, \dots, N$ and $t = 1, \dots, n$,

$$G_{x,i} = H_i\mu_i G_{y,j} = A_j\mu_j H_i = \frac{\sum_{i=1}^{n_i} X_i}{\sum_{i=1}^{n_i} Y_i}, A_j = \frac{\sum_{j=1}^{n_j} X_j}{\sum_{j=1}^{n_j} Y_j}$$

H_i and A_j are the home and away multiplication factors respectively,and μ_i and μ_j are the means of goals scored for team i and j.

Where $\sum_{i=1}^{n_i} X_i$ and $\sum_{i=1}^{n_i} Y_i$ is the total number of home-ground and away-ground goals scored by team i, and $\sum_{j=1}^{n_j} X_j$ and $\sum_{j=1}^{n_j} Y_j$ is the total number of home-ground and away-ground goals scored by team j. Then:

$$d_{it} = \begin{cases} 1 \text{ if home} - \text{ground effect is considered} \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases}$$

The home ground advantage is not considered if the match is considered as a derby, that is, a match between teams from the same town or city and when a team is playing against a relatively weaker team at the weaker team's home-ground.

3.7. Dependence between Home and Away Scores

The dependence coefficient γbetween the two scores in a match is the same for all matches played. A covariance between the home and away goals is calculated each season. A statistical distribution assumed by these covariance coefficients is obtained and the dependence coefficient is estimated through the Monte Carlo maximum likelihood estimation.

3.8. Specific Outcome Probabilities

The goal scoring intensities for the home and away teams, $\lambda_{x,ijt}$, and $\lambda_{y,ijt}$ are used together with γ , the dependence parameter, in the bivariate Poisson probabilitydensity function specified in [1] to compute probabilities for various combinations of home team and away team scores X and Y respectively. The scores X and Y are obtained from the vectors

$$X_w = [1; 2; 2; 3; 3; 3; 4; 4; 4; 4; 5; 5; 5; 5; 5]^T$$

$$Y_l = [0; 0; 1; 0; 1; 2; 0; 1; 2; 3; 0; 1; 2; 3; 4]^T.$$

For a home-team win, the scores are read as combinations of corresponding entries in these given vectors, for example (1;0), (2;0), (2;1), and so on. The number of goals are truncated to5 because the probability of a team scoring more than 5 goals in a match has been discovered to be very small, that is, closer to zero and can be neglected.

For a draw, we compute the probabilities using a combination of vectors

$$X_d = [0; 1; 2; 3; 4; 5]^T$$

$$Y_d = [0; 1; 2; 3; 4; 5]^T.$$

If we assume that the maximum goals a team can score is 5, we have only 6 combinations of scores for draws, i.e. (0 - 0), (1 - 1), (2 - 2), (3 - 3), (4 - 4) and (5 - 5) against30 combinations for a home-team win and an away-team win. This explains the reason why there is always low probabilities for draws compared to wins.

Probabilities for an away-team win are computed using the vectors

$$X_l = [0; 0; 1; 0; 1; 2; 0; 1; 2; 3; 0; 1; 2; 3; 4]^T$$

$$Y_w = [1; 2; 2; 3; 3; 3; 4; 4; 4; 4; 5; 5; 5; 5; 5]^T.$$

3.9. Betting Strategy

A value betting strategy in this study was used. This is when the probability implied by the bookmaker or betting house is less than the true probability, which in this case is the probability according to our model, supposing it gives robust results. To obtain optimum value, we will build a 5% margin of error in our probabilities so that we are 95% certain that our probabilities are robust.

Match outcomes on which a bet is made can be represented by the following decision rule:

Bet \$B when,

$$1 \leq P_{ij}/\Phi_{ij} \leq V \quad [10]$$

P_{ij} is the model predicted probability for our selected

match outcome for a match between team I and team j,Φ_{ij} is the probability implied by the available odds on an outcome for a match between team i and team j,V>1 is the threshold for our probability ratios and should be selected carefully. A big value of V shows a great difference between the model's predicted probabilities and the bookmaker's implied probability. Further analysis should be carried out in such a situation or the gambler should use his or her expertise or intuition.

The strategy is to identify all the matches in a particular week that meet our decision rule and then place bets.

3.10. Specification and Diagnostic Checking

After selecting a model, the next step is its specification. The process of specifying a forecasting model involves selecting the variables to be included, selecting the form of the equation of relationship and estimating the values of the parameters in the equation.

After the model is specified, its performance characteristics should be verified or validated by comparison of its forecasts with historical data for the process it was designed to forecast. Measures such as Mean Absolute Percentage Error (MAPE), Relative Absolute Error (RAE) and Mean Square Error (MSE) maybe used for validating the model. Selection of an error measure has an important effect on the conclusions about which of a set of forecasting methods is most accurate.

Time-series forecasting assumes that a time series is a combination of a pattern and some random error. The goal is to separate the pattern from the error by understanding the pattern's trend, its long-term increase or decrease, and its seasonality, the change caused by seasonal factors such as fluctuations in use and demand.

In diagnostic checking we check whether the model represents accurately the underlying process in the given time series. The following tests were performed:

3.11. Test for Constant Variance

The test analyses the trend in the residuals and makes sure that the assumption of constant variance is not violated. Analysis is done using data plots of residuals versus time. The Durbin-Watson test is performed to check for model adequacy under the hypothesis:

H_0: There is no autocorrelation between residuals (d = 2)

H_0: There is autocorrelation between residuals ($d \neq 2$),

where,

$d = \frac{\sum_{t=2}^{n}(U_t - U_{t-1})^2}{\sum_{t=1}^{n} U_t^2}$ and U_t and U_{t-1} the error terms.

3.12. Test for Independence

From the sample, the ACF of residuals is computed. Residuals are independent if they do not form any patterns on the ACF plot. From the ACF plot, if all the autocorrelations are insignificant then the residuals are independent. The runs test is performed using Minitab to test for independence of residuals under the

Hypothesis:

H_0: The residuals are not dependent on each other.

H_0: The residuals are dependent on each other.

3.13. Test for Normality

Testing for normality is done by plotting a histogram of residuals. Histogram of normally distributed residuals should approximately be symmetric and bell shaped.

A normal score test is done by plotting the residuals against normal score. With normally distributed data, the plot of the ordered data values versus the corresponding normal scores should fall approximately on a straight line.

4. Research Findings and Simulations

A time series analysis of ten years of football match results from the Zimbabwe Premier Soccer League for which 16 football clubs are active in each season was done. The 16 football clubs that participate in a season vary because the four lowest placed teams at the end of the season are relegated. In the new season, they are replaced by four other teams. The number of different teams in the panel is 40 and the data for all the 40 football clubs were captured. For predictions and implementation of the betting strategy the focus will be shifted to only 16 teams that were part of the 2013 season. CAPS United FC, Dynamos FC, Highlanders FC and Motor Action FC are the only teams that have played in all ten seasons of the sample while ten other teams have only played in one season due to relegation or promotion.

4.1. Distribution of Goals

The mean number of goals scored is 1.16 and the Poisson distribution models the data.

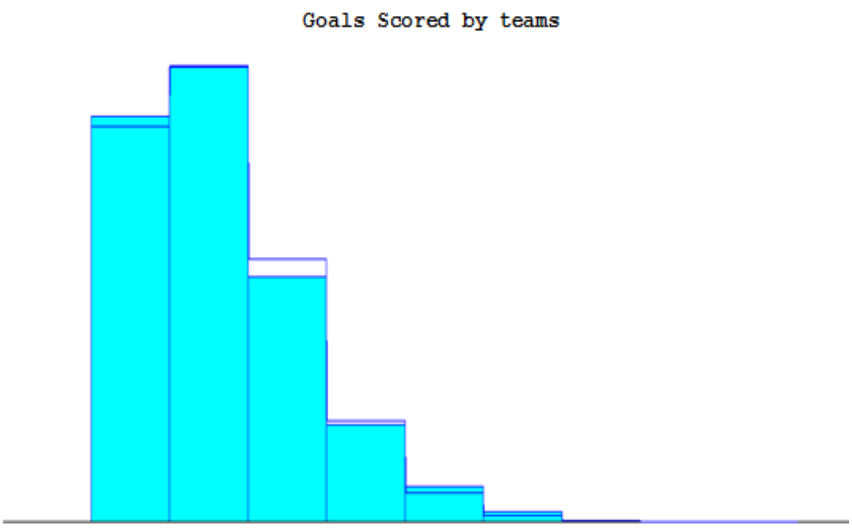

Figure 1. *Distribution of goals.*

Since$< \alpha$, that is, 0.005<0.05, we conclude that the goals distribution follow a Poisson distribution at 5% level of significance. This validates the assumption that the scores by the home and away teams are generated from a Poisson distribution.

4.2. Home and Away Team Performance

To determine whether or not the home-ground performance of a team is better than the away-ground performance, difference of two means testing was done. To test whether the mean of home scores is significantly greater than the mean of the away scores at $\alpha = 0.05$ level of significance, a two-sample t-test is carried out and the null hypothesis is rejected since the computed value of the test statistic $t_0 = 1.68 > t_{0.05,78} = 1.66$, the critical value. The mean of home scores is significantly greater than the mean of away scores.

4.3. Home Ground Advantage

To test whether different teams in the league have different home-ground advantages or not, an ANOVA was developed for the fixed-effects and the null hypothesis is rejected since the test statistic $F_0 = 1.89 > F_{crit} = 1.50$ and conclude that the treatments or factor level means are different.

Carrying out the multiple comparisons using the Fisher's Least Significant Difference (LSD) method to determine which means are significantly different, it was found that teams have different home-ground effects.

4.4. Attack and Defence Capabilities

A time series plots of the number of goals scored from the 2004to the 2013 season by the 2 high ranking and 2 low ranking teams as per the end of the 2013 season was done.

The high ranking teams are Dynamos and Highlanders and the low ranking teams are Monomotapa and Motor Action. All the teams with the exception of Monomotapa appear in all the ten seasons. Monomotapa does not appear in the 2004 season. A time series plots of the number of goals conceded from the 2004 to the2013 season by the 2 high ranking and 2 low ranking teams as per the end of the 2013 season was also done.

From the panel data set, 16 attack capability coefficients and 16 defence capability coefficients for the teams in the 2013 season need to be estimated. If all teams play their matches on a weekly basis, each season consists of 30 weeks.

We therefore obtain our attack and defence capabilities as pth - order autoregressive time series forecasts of goals scored for attack capability, α_i and goals conceded for defence capability β_i. The value of p is team dependent, that is, from the analysis, other teams' goals are auto-regressive processes of order 1 and others are of order 2. The order is obtained from ACF and PACF plots.

4.5. The Dependence Coefficient

The dependence in the scoring intensities of two opposing teams isγ. According to Karlis and Ntsoufras (2003), a value of γas low as 0.05, has a significant impact on the model results. The researchers point out that Poisson models tend to underestimate the probability of draws. The dependence parameter when covariance coefficients between the home and away scores over the years are calculated. These covariance coefficients are uniformly distributed over the years; therefore a Monte Carlo maximum likelihood estimation of this parameter is performed. Table1 gives the Monte Carlo estimates of the dependence coefficient.

Table 1. *Dependence parameter estimation.*

Parameter	500 repetitions	1 000 repetitions	2 000 repetitions
γ	0.0555	0.0553	0.0551
SD	0.0286	0.0294	0.0294

A total of 500, 1000 and 2000 repetitions of the simulation are done and a value of $\gamma = 0.055$ is obtained. This low correlation between scores has an impact on the probability of draws and so a bivariate Poisson distribution cannot be ignored.

4.6. Other Parameter Estimates

The home-ground coefficient and other model parameters are estimated from the equations given in Section 3 and are presented in Table 2. The defence capability coefficients are assigned negative values as they are the goals a team expects to concede.

Table 2. *Model parameters.*

i	Team	μ_i	H_i	A_i	δ_i	$G_{x,i}$	$G_{y,i}$	α_i	β_i
1	B Mambas	1.09	1.81	0.65	0.27	1.97	0.70	1.22	-1.19
2	B Rhinos	1.15	1.13	0.97	0.02	1.30	1.12	0.45	-1.06
3	Buffaloes	1.08	1.71	0.68	0.28	1.85	0.73	1.04	-1.45
4	CAPS United	1.32	1.39	0.77	0.19	1.84	1.02	1.23	-1.02
5	Chicken Inn	1.35	1.35	0.74	0.19	1.83	1.00	0.70	-1.22
6	Dynamos	1.38	1.41	0.71	0.16	1.95	0.98	1.13	-0.75
7	FC Platinum	1.33	1.22	0.87	0.11	1.63	1.16	0.54	-0.94
8	Harare city	1.08	1.61	0.62	0.25	1.75	0.67	0.58	-1.17
9	Highlanders	1.36	1.41	0.65	0.23	1.92	0.88	1.18	-0.89
10	Howmine	0.90	0.93	1.08	-0.03	0.84	0.97	1.48	-0.90
11	Hwange	1.20	1.85	0.60	0.33	2.23	0.72	0.36	-0.42
12	Monomotapa	1.18	1.34	0.88	0.11	1.59	1.04	1.14	-0.83
13	Motor Action	1.21	1.17	0.96	0.09	1.42	1.16	1.19	-0.86
14	Shabanie Mine	1.24	1.67	0.69	0.24	2.07	0.85	1.03	-1.66
15	Triangle	1.43	1.39	0.72	0.23	1.99	1.03	1.10	-0.98
16	Tripple B	0.87	2.25	0.44	0.33	1.95	0.39	0.27	-1.62

4.7. Model Validation

The bivariate Poisson model relies on having accurate goal scoring intensities for each team. A goal scoring intensity represents the number of goals team *i* is expected to score against team *j*. These goal scoring intensities, to be estimated inthe next section, rely on accurate estimates of attack and defence capabilities. Model validation based on the estimates of these parameters was conducted.

The following diagnostic checks were conducted to check whether the model represents accurately the underlying process in the given time series on Dynamos, Highlanders, Monomotapa United and Motor Action.

4.8. Runs Test

Runs tests for the four teams selected were conducted. The null hypothesis (H_0) was not rejected at 5% level of significance and concluded that the sequence was produced in a random manner.

4.8.1. Autocorrelation of Residuals

If the residuals u_t are 'white noise' then the ACF of $u_t \sim N(0\,;\, 1/\sqrt{n})$, thus spikes are significant if they are outside the range $\pm 2/\sqrt{270}$. From the ACF of residuals, it is noted that the residuals are uncorrelated since there are no significant spikes.

4.8.2. Normality Tests

Normality is investigated through the use of histograms and normal plots of residuals. Figure 2 shows the normality charts for the four selected teams.

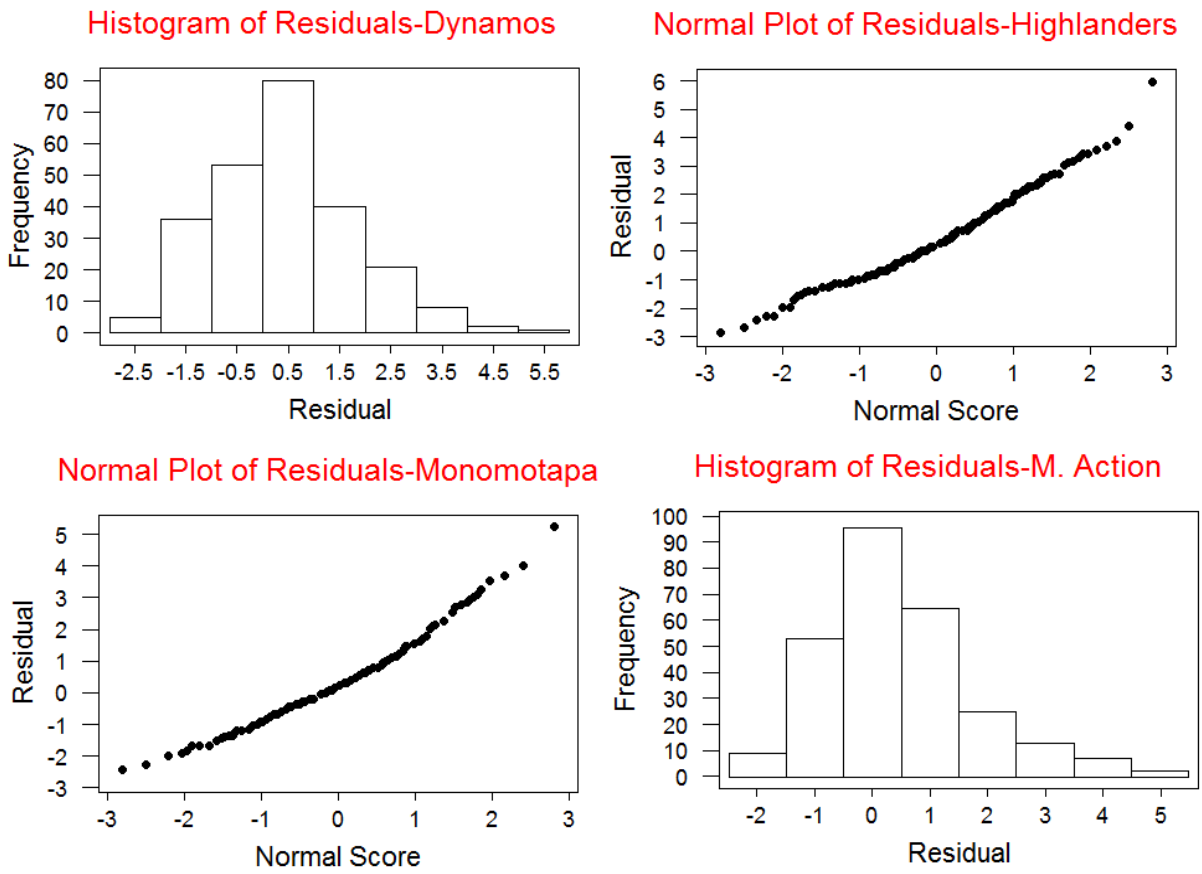

Figure 2. *Normality plots.*

From the normal plots and histograms of residuals in Figure 2, the normality assumption was not violated. However, the histograms of residuals indicate a slight skewness.

4.8.3. Randomness of Residuals

Figure 3 below is a plot of residuals against fits.

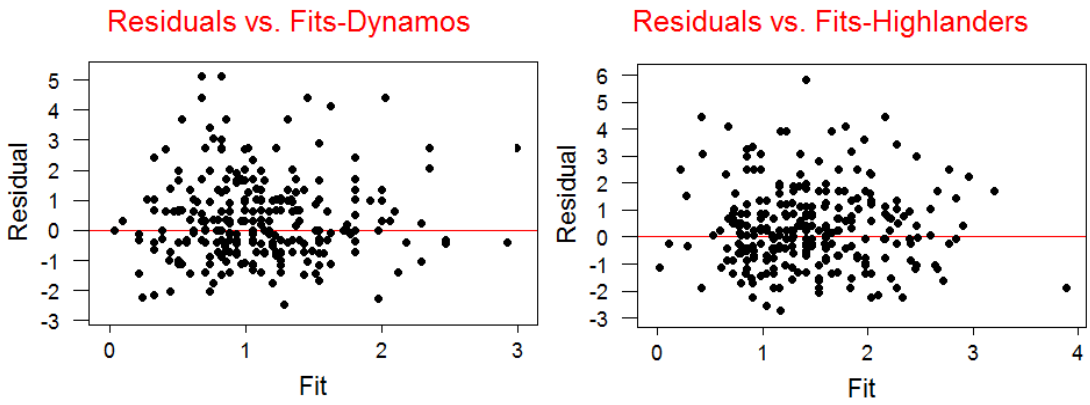

Figure 3. *Residuals versus fits.*

The plot of residuals against fits does not exhibit any pattern whatsoever, therefore the residuals are random. From these tests, one can conclude that the residuals approximate a normal distribution with mean zero and are random.

4.9. Goal Scoring Intensities

The model parameters given in Table 3 are used to obtain scoring intensities λ_x and $\lambda_y$as per equation [9] and shown in Table 3. Each match has a unique set of scoring intensities and these are estimated for each soccer match and calculated after updating the previous scores and estimating new defence and attack capabilities for the teams. These scoring intensities therefore vary with teams and time, they are dynamic.

Table 3. *Goal Scoring Intensities.*

WEEK 1	λ_x	λ_y	WEEK 16	λ_x	λ_y	WEEK 30	λ_x	λ_y
Match			**Match**			**Match**		
1	0.72	1.06	120	0.24	0.85	232	1.56	0.74
2	0.87	0.78	121	2.68	0.91	233	1.21	1.39
3	1.38	0.70	122	1.80	1.25	234	1.24	0.75
4	1.25	0.94	123	2.04	0.56	235	0.63	1.13
5	0.63	1.08	124	0.58	0.64	236	0.56	0.87
6	1.35	2.03	125	0.54	0.55	237	0.27	1.08
7	0.83	0.88	126	1.90	1.34	238	1.36	0.66
8	2.19	1.66	127	2.12	1.42	239	0.70	2.44

4.10. Match Outcome Prediction

The model presented in [1] is used to determine the specific match outcome probabilities and are summarised in Table 4 below. Due to the time-consuming computation of these individual outcome probabilities, we limit our analysis to three weeks of the 2013 football season; week 1 (beginning of the football season), week 16 (midseason)and week 30 (end of season).

Table 4. *Outcome probabilities.*

Match Number	Home team	Away team	P_H	P_D	P_A
	WEEK 1				
1	Harare City	CAPS United	0.245	0.324	0.430
2	How Mine	Triple B	0.352	0.346	0.301
3	Motor Action	Hwange	0.529	0.278	0.189
4	Black Mambas	Dynamos	0.432	0.287	0.281
5	Black Rhinos	Monomotapa United	0.212	0.327	0.461
6	Buffaloes	FC Platinum	0.250	0.214	0.535
7	Chicken Inn	Triangle	0.316	0.339	0.345
8	Shabanie Mine	Highlanders	0.698	0.177	0.125
	WEEK 16				
120	Chicken Inn	Shabanie Mine	0.100	0.408	0.491
121	Black Mambas	Triangle	0.254	0.149	0.597
122	How Mine	Dynamos	0.491	0.231	0.278
123	Harare City	Hwange	0.702	0.187	0.111
124	Buffaloes	Triple B	0.595	0.250	0.156
125	Black Rhinos	FC Platinum	0.275	0.444	0.281
126	Motor Action	Monomotapa United	0.491	0.223	0.286
127	Highlanders	CAPS United	0.512	0.209	0.279
	WEEK 30				
232	CAPS United	Harare City	0.371	0.237	0.379
233	Triple B	How Mine	0.294	0.349	0.357
234	Hwange	Motor Action	0.429	0.235	0.322
235	Dynamos	Black Mambas	0.536	0.249	0.207
236	Monomotapa United	Black Rhinos	0.075	0.225	0.692
237	FC Platinum	Buffaloes	0.405	0.278	0.313
238	Triangle	Chicken Inn	0.477	0.297	0.224
239	Highlanders	Shabanie Mine	0.482	0.246	0.263

In Match 6 in the first week, Buffaloes is playing at home ground against FC Platinum. The probability for a Buffaloes win is 0.250, probability for a draw is 0.214 and the probability for a Platinum win is 0.535. Therefore Platinum

are the favourites in this match. The outcomes with the highest probabilities are therefore our required results, for instance, in match 122, which is in week 16, we expect Howmine FC to beat Dynamos FC.

4.11. Betting Strategy

To implement our betting strategy defined in Section 3, forweek 30 of the 2013 soccer season, we use the 8 matches played on the 24thof November 2013 to test our betting strategy. These are matches 232 - 239 shown in Tables 5. The betting selections are based on outcomes with the highest probabilities. The team selected is the one a bet is placed on to win the match. We multiply our model outcome probabilities by 0.95 so as to factor in a 5% error margin. The rule in [10] is used in deciding whether to bet or not to bet.

Table 5. *Betting Strategy.*

Match Number	Bet Selection	Model Probability	Bookmaker's Probability	Probability Ratio	Decision
232	Harare City	0.360	0.324	1.1	Bet
233	How Mine	0.339	0.421	0.8	No Bet
234	Hwange	0.408	0.382	1.1	Bet
235	Dynamos	0.509	0.383	1.3	Bet
236	Black Rhinos	0.657	0.346	1.9	Bet
237	FC Platinum	0.384	0.241	1.6	Bet
238	Triangle	0.453	0.366	1.2	Bet
239	Highlanders	0.458	0.523	0.9	No Bet

As per our betting strategy, there are six matches to lay bets on and these are indicated in Table 5. Supposing we have a $100 amount for betting, for the six matches we make separate bets and split our amounts proportionally in terms of the given odds. The bet amounts and the returns expected are shown in Table 6. The analyses show that, the total net return is $169.52 provided that all the individual bets are won. However, this net return is usually higher in practice because different bookmakers use various multiplicative factors to arrive at the final pay-out. In matches 232 and 237, bets were lost since the respective teams failed to win the matches. Therefore our net return becomes $109.03. If on the other hand a single bet for all the matches was made for all the matches, $100 would have been lost.

Table 6. *Betting Returns.*

Match Number	Bet Selection	Odds	Implied Probability	Bet Allocation	Net Returns	Bet Results
232	Harare City	2/1	0.33	$15	$29.95	Lose
234	Hwange	15/8	0.35	$16	$29.30	Win
235	Dynamos	11/9	0.45	$20	$24.71	Win
236	Black Rhinos	3/2	0.40	$18	$26.95	Win
237	FC Platinum	17/8	0.32	$14	$30.55	Lose
239	Highlanders	5/3	0.38	$17	$28.08	Win
	TOTAL			$100.00	$169.52	

4.12. Validation Using Existing Data

Scores for the 2013 soccer season are used to validate our model. The actual weeks for which predictions were made are week 1, week 16 and week 30, therefore validation is done using the data from these particular weeks.

4.12.1. Actual Outcomes Versus Model Outcomes

The actual scores are compiled for the respective matches and are shown in Table 7. A comparison is done between actual outcomes and the model outcomes.

Table 7. *Actual outcomes versus model outcomes.*

Match Number	Home team	Away team	X	Y	Actual Result	Model Result
	WEEK 1					
1	Harare City	CAPS United	1	0	Home win	Away win
2	How Mine	Triple B	2	0	Home win	Home win
3	Motor Action	Hwange	0	0	Draw	Home win
4	Black Mambas	Dynamos	1	1	Draw	Home win
5	Black Rhinos	Monomotapa United	3	2	Home win	Away win
6	Buffaloes	FC Platinum	1	2	Away win	Away win
7	Chicken Inn	Triangle	2	1	Home win	Away win
8	Shabanie Mine	Highlanders	0	2	Away win	Home win

Match Number	Home team	Away team	X	Y	Actual Result	Model Result
	WEEK 16					
120	Chicken Inn	Shabanie Mine	1	1	Draw	Away win
121	Black Mambas	Triangle	0	1	Away win	Away win
122	How Mine	Dynamos	1	0	Home win	Home win
123	Harare City	Hwange	2	1	Home win	Home win
124	Buffaloes	Triple B	5	1	Home win	Home win
125	Black Rhinos	FC Platinum	2	0	Home win	Draw
126	Motor Action	Monomotapa United	0	2	Away win	Home win
127	Highlanders	CAPS United	0	0	Draw	Home win
	WEEK 30					
232	CAPS United	Harare City	2	2	Draw	Away win
233	Triple B	How Mine	2	2	Draw	Away win
234	Hwange	Motor Action	1	0	Home win	Home win
235	Dynamos	Black Mambas	2	0	Home win	Home win
236	Monomotapa United	Black Rhinos	0	1	Away win	Away win
237	FC Platinum	Buffaloes	2	2	Draw	Home win
238	Triangle	Chicken Inn	2	1	Home win	Home win
239	Highlanders	Shabanie Mine	3	1	Home win	Home win

In the first week of the season, only 3 matches out of 8 are predicted correctly. This is due to the fact that there are teams new to the league that had an unpredictable performance, very little or no information was available for these teams. Parameter estimates for the end of the soccer season had to be used to predict outcomes for the beginning of the soccer season. For older teams, quite a number of changes could have occurred from the 2012 season, for instance, players could have transferred, new acquisitions occurred, the teams lacked purpose or some did not take the match seriously. In week 16 of the 2013 season 4 out of 8 matches are correctly predicted. Again this may be due to the fact that parameter estimates for the end of the soccer season had to be used to predict these outcomes. The dynamic nature of our model is further emphasized when week 30 matches are predicted. The parameter estimates for the end of the season are used in this case and as a result,5 out of 8 matches are predicted correctly. The other 3 matches had resulted in draws and our model, being a Poisson, failed to pick the draws.

4.12.2. Model Probabilities Versus Actual Probabilities

Actual probabilities are computed from the actual matches played and the actual outcomes in a particular week. The actual probabilities are the actual outcomes over possible outcomes in a given week for a home team win, draw and an away team win. The model probabilities are the average probabilities obtained from the bivariate Poisson model.

Table 8. *Model probabilities versus actual probabilities.*

Model Probability				Actual Probability			
Week	**Home-win**	**Draw**	**Away-win**	**Week**	**Home-win**	**Draw**	**Away-win**
1	0.38	0.29	0.33	1	0.50	0.25	0.25
16	0.43	0.26	0.31	16	0.50	0.25	0.25
30	0.38	0.26	0.34	30	0.57	0.14	0.29

Table 8 shows the probabilities for a home win, draw and an away win from our model and the actual probabilities. To test whether probabilities are significantly different, a two-sample *t*-test, assuming unequal variances, is performed. The test is shown in Table 9.

Table 9. *T-test: Two-Sample Assuming Unequal Variances.*

	Model	Actual
Mean	0.332	0.333
Variance	0.003	0.022
Observations	9	9
Hypothesized Mean Difference	0	
df	10	
t Stat	-0.017	
$P(T \leq t)$ one tail	0.493	
*t*Critical one-tail	1.812	
$P(T \leq t)$ two tail	0.987	
*t*Critical two-tail	2.228	

We fail to reject the null hypothesis and conclude that the actual and model probabilities are not significantly different.

5. Conclusion and Recommendations

The goals scored by teams in soccer were found to follow Poisson distributions. A relatively low dependence coefficient was obtained and the interpretation is that the goals by any two teams playing against each other are correlated. The research confirmed that the home ground advantage exists in the Zimbabwe Premier Soccer league. It was shown that this effect varies with teams; older teams usually have a greater advantage compared to teams new to the league. A value betting strategy was devised and was noted to be profitable in the longrun. In addition to their ad-hoc techniques, a statistically proper system that predicts probabilities with reasonable accuracy and also monitors the betting distribution is a vital tool for the bookmakers as more and more matches need to be covered and more attractive prices need to be compiled.

References

[1] Bittner, E., NuBbaumer, A., Janke, W. and Weigel, M. (2008). Football fever: goal distributions and non-Gaussian statistics. Eur. Phys. J. B 67, 459471

[2] Bulla, J., Chesneau, C., Kachour, M. (2012). On the bivariate Skellam distribution. Journal of Multivariate Analysis.

[3] Cengiz, M. A., Murat, N., Koc, H., Senel, T. (2012). A Bayesian Computation for the Prediction of Football Match Results Using Artificial Neural Network. International Journal of Scientific Knowledge (Computing and Information Technology) Volume 1 Issue 2. [Online] Available from: www.ijsk.org. [Accessed on18/08/13]

[4] Crowder, M., Dixon, M., Ledford, A. and Robinson, M. (2002). Dynamic modeling and prediction of English Football League matches for betting. The Statistician51, 157-168.

[5] Dixon, M. J. and Coles, S. C. (1997). Modelling association football scores and inefficiencies in the football betting market. Applied Statistics 46, 265-280.

[6] Goddard, J. and Asimakopoulos, I. (2004). Forecasting football results and the efficiency of fixed-odds betting. J. Forecasting 23, 51-66.

[7] Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data using bivariate Poisson models. The Statistician 52, 381-393.

[8] Karlis, D. and Ntzoufras, I. (2009). Bayesian modelling of football outcomes: using the Skellam's distribution for goal difference, IMA J. Numerical Analysis 20, 133-145.

[9] Koopman, S. J. and Lit, R. (2012). A Dynamic Bivariate Poisson Model for Analysing and Forecasting Match Results in the English Premier League. Tinbergen Institute Discussion Paper.

[10] Lotteries and Gaming Act [Chapter 10:26]. 2000. 109-118.

[11] Oberhofer, h., Philippovich, T. and Winner, H. (2009). Distance matters in awaygames: Evidence from the German Football League. University of Salzburg. Workingpaper No. 2009 - 01.

[12] O`Shaughnessy, D. (2012). Optimal exchange betting strategy for win-draw-loss markets. Ranking Software, Melbourne.

[13] Pollard, R. (2006). Home advantage in soccer: variations in its magnitude and a literature review of the inter-related factors associated with its existence. J. Sport Behavior 29, 169-189.

[14] Rue, H. and Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league, J. Royal Statistical Society D 49, 399-418.

[15] Saavedra, G. M., Gutierrez, A. O., Fernandez, R.J.J. and Marques, P. (2012). Measuring home advantage in Spanish football. Sport training. Council of Europe classification: 17. Metrology of Sport.

[16] Vlastakis, N., G. Dotsis, and R. N. Markellos (2007). How Efficient is the European Football Betting Market? Evidence from Arbitrage and Trading Strategies. Journal of Forecasting.

[17] ZIFA(2013). www.zifa.co.zw